Intelligent Instructional Hand Offs

Stephen E. Fancsali
Carnegie Learning, Inc.
501 Grant Street, Suite 1075
Pittsburgh, PA 15219, USA
+1 (412) 992-5099
sfancsali@
carnegielearning.com

Michael V. Yudelson
ACT, Inc./ACT Next
500 ACT Drive
Iowa City, IA 52243, USA
michael.yudelson@act.org

Susan R. Berman
Steven Ritter
Carnegie Learning, Inc.
501 Grant Street, Suite 1075
Pittsburgh, PA, 15219, USA
{sberman, sritter}@
carnegielearning.com

ABSTRACT

Learners in various contemporary settings (e.g., K-12 classrooms, online courses, professional/vocational training) find themselves in situations in which they have access to multiple technology-based learning platforms and often one or more non-technological resources (e.g., human instructors or on-demand human tutors). Instructors, similarly, find themselves in situations in which they can provide learners with a variety of options for instruction, practice, homework, and other activities. We seek data-driven guidance to help facilitate intelligent instructional "hand offs" between learning resources. To begin this work, we focus on an important element of self-regulated learning, namely help seeking. We build classifier models based on proxies for learner prior knowledge and data-driven inferences about learners' disengaged behavior (e.g., gaming the system) and affective states (e.g., confusion) to determine the extent to which (and when) learners tended to seek out help via human tutoring while using an intelligent tutoring system for mathematics. Insights into cognitive, behavioral, and affective factors associated with help seeking outside of a system will drive future work into providing automated, intelligent guidance to both learners and instructors. We close with discussion of the limitations of the present analysis and avenues for future work on intelligently guiding instructional hand offs.

Keywords

intelligent tutoring systems, Cognitive Tutor, mathematics education, developmental mathematics, higher education, online courses, human tutoring, detector models

1. INTRODUCTION

The proliferation of technology-based learning platforms and applications (apps), including intelligent tutoring systems (ITSs), game-based learning environments, massively open online courses (MOOCs), training simulators, language learning apps, and practice apps, among others, creates a complex array of choices for learners and those who would seek to facilitate learning. Far from replacing human instruction, these technologies are often used in learning environments in which learners have access to both technological and human sources of instruction.[1]

Instead of comparing the relative effectiveness of technological and human instruction (c.f. [12, 34]), we are concerned with the extent to which learners' interactions with both technology-based and human resources can be treated as a system that is a target for optimization. One key target for optimizing such a system is the ability to intelligently guide "hand offs" or transitions between different learning apps and to guide learner help seeking as they use technology but also have access to (limited) human resources like an instructor or tutor. Work that considers such hand offs, and intelligent guidance for them (e.g., when a system or app could best provide feedback that directs the learner to an external resource because they need help or could benefit from practice on a pre-requisite skill that is not covered by the system or app), is limited, though one noteworthy exception attempts to provide adaptive assistance as students learn to program by suggesting open, online reading content related to errors made while the student programs [33].

One key element of self-regulated learning [37] is the ability for learners to appropriately and effectively seek out and use help when they need it [3, 27]. ITSs and other technology platforms for learning frequently provide learners with hints and other forms of scaffolding, guidance, and help. Unfortunately, learners often do not make efficient or extensive use of such help within ITSs [1, 25, 36], and when they do, learners sometimes "abuse" such help [2], whether by rapidly seeking progressively more informative hints or attempting to "game the system" [6]. More recent work begins to explore when students *ought* to seek help *within* an ITS. For example, one study found that help avoidance earlier in the problem solving sequence, as students solve genetics problems in an ITS for genetics, is more strongly and negatively associated with robust learning outcomes, suggesting that early help seeking ought to be encouraged [4]. Work like that of [4] is a part of a broader literature focusing on providing meta-cognitive support and developing "meta-cognitive" tutors (e.g., [2]).

Classroom practices in blended, K-12 classrooms also encourage self-regulated learning. Here, students typically have direct access to a teacher while they work within an environment like an ITS. Teachers often adopt strategies like "ask three then me" [17] to

[1] The second author's primary contribution to this work was made while he was employed by Carnegie Learning, Inc., and later Carnegie Mellon University.

encourage productive behavior with respect to help seeking, rather than over-reliance on the teacher. Following this strategy, for example, the student might use the hint feature of an ITS, and should that not provide sufficient clarity or guidance, ask the student on each side of her in the classroom before asking the teacher for help. Given tendencies to over-use and under-use help, better student self-regulation is one important element in optimizing the teacher's scarce time. Ideally, over-users of help will start to rely on help provided by the ITS or their peers, encouraging productive collaboration among learners and enabling teachers to spend more time with students experiencing genuine struggle with content or who rarely seek out help despite needing it.

In the present study, rather than a traditional or blended K-12 classroom, we consider use of the Cognitive Tutor [26] ITS, in one or more of a sequence of two, five-week, fully online developmental mathematics course at a large, mostly-online university. In addition to an instructor, available to students via e-mail and an online message board, students in these courses had optional and unlimited access to human mathematics tutors via a service called Tutor.com (TDC). We were able to obtain access to all chat logs with TDC, as well as detailed data on CT use, providing an ideal dataset to investigate how students navigated between human and automated support in this environment.

In the present study, we focus on cognitive, behavioral, and affective factors that predict whether (and the extent to which) students using CT seek out help from human tutors via an online chat service called Tutor.com. To do so, we adopt a discovery with models approach [10] and build classifier models based on proxies for learner prior knowledge and data-driven inferences about learners' disengaged behavior (e.g., gaming the system, guessing, off-task behavior) and affective states (e.g., confusion, boredom), relying on "detector" models of such factors [5-9]. Insights into cognitive, behavioral, and affective factors associated with help seeking outside of an ITS will drive future work into providing automated, intelligent guidance to both learners and instructors.

2. COGNITIVE TUTOR (CT) & TUTOR.COM (TDC)

Cognitive Tutor (now called MATHia in K-12 contexts and Mika in higher education contexts) is a mathematics ITS developed and distributed by Carnegie Learning, Inc. [26], used by hundreds of thousands of learners each year in K-12 and higher education learning contexts (see Figure 1).

As illustrated in Figure 1, learners in CT work through complex, multi-step math problems. Within each problem, steps are mapped to fine-grained skills or knowledge components (KCs) [24]. KC mastery is tracked using Bayesian Knowledge Tracing [15].

CT's instructional approach is based on mastery learning [11], and it relies on BKT and these parameters to update estimates of a learner's mastery of the KCs it tracks, as they practice and learn the KCs, within each of its topical sections of content. Within each section, CT presents problems to learners that emphasize the KCs they have yet to master. After mastering all KCs in a section, learners "graduate" to the next section. Having failed to master all of a section's KCs by a certain pre-set limit (e.g., a maximum number of problems), the learner is "promoted" to the following section. MATHia/Mika analytics provide the teacher with information about graduation and promotion status; in promotion cases, teachers will know that the student has failed to master KCs for a particular topic, allowing them to provide some form of remediation, including possibly allowing for a second attempt to work through problems in the ITS later.

As students learn and practice, CT provides context-sensitive, adaptive hints and other feedback. In a typical, blended, K-12 classroom environment in which CT is frequently used, students using CT are in physical proximity to their fellow students and teachers, so they can rely on these resources for help if, for some reason, the CT is not providing sufficient feedback and help. In the present context, CT is used in a fully online context, so for real-time help, the student has to rely on human math tutors, made available to them via an online chat mechanism provided by Tutor.com (TDC). Student could also communicate asynchronously with their course instructors via e-mail and with their fellow students and instructor via an online message board, but data surrounding these means of communication were unavailable to the authors.

TDC is a large provider of online, one-to-one, and on demand tutoring for students in a variety of domains and settings (including learners in K-12 public schools, colleges, universities, libraries, corporations, and the U.S. military). In the context of the present study, TDC tutors were accessible to students, via an online chat mechanism, as a part of their enrollment in the two developmental math courses of which CT was a mandatory instructional component and the primary means by which students were provided with problem-solving practice and exercises. Students were typically assigned several units of content (i.e., sets of sections of content) for each week of the course and allowed, generally, to progress at their own pace through those sections with the expectation that they would complete assigned content within the week in which it was assigned or shortly thereafter.

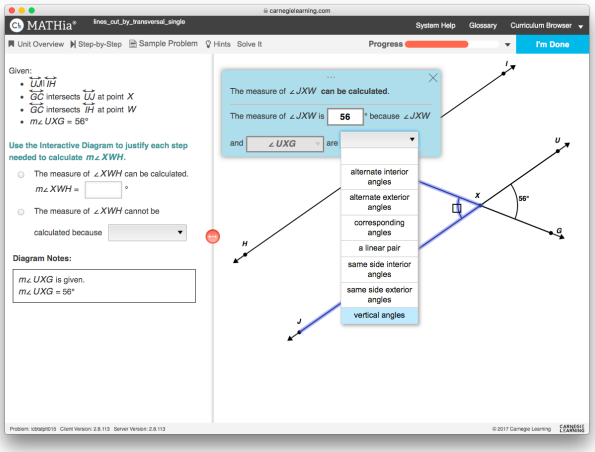

Figure 1. Cognitive Tutor/MATHia/Mika screenshot

3. DATA

For the present study, the population of concern is comprised of 16,905 adult students in at least one course (and in many cases both courses) in a sequence of two, five-week development mathematics courses for the time period of June 2014 to December 2014, inclusive. Of these 16,905 students over this time period, 80.4% (13,585) made no use of TDC. 3,320 students used TDC at least once during at least one of these courses, with a total number of 19,248 TDC sessions taking place over this six-month period. Tutoring chat sessions lasted from several minutes to over an hour, many occurring while learners simultaneously used CT

Though outside the scope of this study, data available also included transcripts of the TDC tutoring chat sessions (and annotations of dialogue acts [32], instructional modes, and switches between these modes within these chat sessions) that allows for sophisticated analyses of interactions between human and automated tutoring systems like ITSs. These topics, using data from this context, have been explored elsewhere [28-29]. However, data like demographics, student background, and performance in other courses were not available to the authors.

In the analysis that follows, we consider a subset of this population, including 3,119 students who used TDC at least once (i.e., all of the students for whom data could be processed for analysis) as well as a random sample of 1,874 students who did not use TDC over this time period.[2] For these students, we have extensive usage data from CT and rely on the timestamps at which TDC sessions started to identify, for example, the CT login session that occurred before each TDC session. We also know, for each TDC User, the number of times they accessed TDC tutoring sessions as well as the duration of these sessions.

CT data for these 4,993 students were processed into a format amenable to the LearnLab DataShop [20, 23]. These data are comprised of 88,497,091 learner actions (i.e., attempts at steps within problems, or tutor transactions in the DataShop parlance) (an average of 17,724 tutor transactions per student).

The second course in the two-course sequence was more advanced and contains both more challenging content (as measured by CT hints requested and errors made) and fewer sections than the first course in the sequence. Nevertheless, there appear to be few major differences in TDC usage (considering session counts, etc.) between the two courses, so our analyses combine data from the two courses. However, not every student in the sample considered was enrolled in both courses over the time window we consider, so some students only have usage data from the first course and some only from the second course.

4. INITIAL OBSERVATIONS & RESEARCH QUESTIONS

Two related, initial observations inform the analyses of the rest of this work. The first relates to the extent to which a small minority of users accounts for a majority of TDC use. The second observation concerns the imbalance in the data, which informs the overall analytic approach we adopt.

4.1 TDC Super-Users, TDC Users, and TDC Non-Users

Figure 2 provides a histogram counting the number of students with a particular number of TDC sessions. As noted previously, over 13,000 students make no use of TDC and have zero TDC sessions. However, the long right tail of this histogram points to a small minority of students who have tens or even hundreds of TDC sessions. We call students in the top 10% of TDC usage (by session, considering only students with at least one session) "TDC Super-Users." The set of TDC Super-Users is comprised of 350 students (or 2.1% of students in these courses over this period) and account for 55.4% of total TDC session time (4,100 hours of TDC session time). On average, TDC Super-Users spent 7.6 hours in TDC sessions over the period of one of these courses. TDC Users (2,769 students with at least one TDC session but who are not TDC Super-Users) spent a total of 3,367 hours in TDC sessions over this time period, with an average of .8 hours of TDC session time per course.

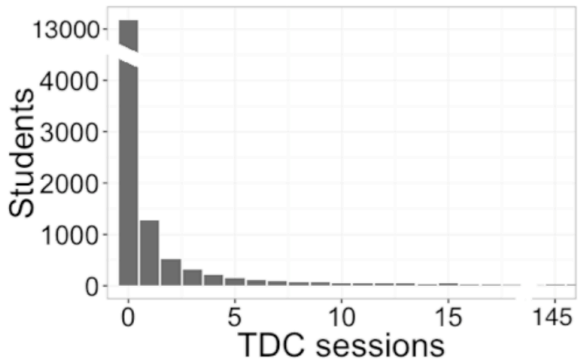

Figure 2. Histogram of TDC sessions and student counts over both courses in the two-course developmental math sequence. Reproduced from [Fancsali, et al unpublished report].

Perhaps unsurprisingly, TDC Super-Users also spent more overall time in CT with an average of 61.7 hours of CT time per course. TDC Users spent an average of 48 hours in CT per course while TDC Non-Users spent only an average of 29.3 hours per course in CT. A more extensive analysis of specific differences and comparisons on various performance metrics for these groups within CT is found in [21].

Such numbers seem likely, though not necessarily[3], to reflect over-use and near-certain under-use (for TDC Non-Users) of the human tutoring provided by TDC. Such over-use and under-use could reflect an underlying problem in terms of self-regulated help seeking. As such, two research questions are directed at the possibility of predicting whether a student is likely to be a TDC Super-User or a TDC User (versus TDC Non-Users). What are possible drivers of such extensive use of TDC? What behaviors and affective states might indicate a need for external help? At a more granular level, the third question seeks to determine whether it is possible to predict from data from a particular CT login session that a student is likely to seek out TDC.

As noted earlier, we center our attention on cognitive features (related to prior preparation for the course), behavioral features like gaming the system, and affective features like boredom,

[2] Seemingly arbitrary counts of 3,119 students who used TDC at least once and the random sample of 1,874 students who never used TDC are largely the result of data collection and data processing limitations in the legacy deployment of Cognitive Tutor used by these students. Some students' data were not reliably collected and/or processed (leading to the difference between 3,320 TDC Users and 3,119 students considered), and time constraints made it impossible to consider a larger sample of TDC Non-Users. Fortunately, present-day implementations of MATHia and Mika no longer suffer from such limitations.

[3] TDC Super-Users, the set of which, for example, could include learners with some form of learning disability like dyscalculia, may derive great learning benefit from interacting with these tutors at the level at which they do (and may need relatively intense remediation to succeed), but this benefit comes at the relatively greater expense of the real-time, chat-based tutor, compared to, for example, regularly setting up time to interact with the instructor or finding other resources for the student to consider when they need such intense help.

among others, that may help to inform future work and provide practical guidance to teachers and facilitators of instruction.

4.2 Research Questions

For each of the following questions, there is a corresponding prediction task for which we consider cognitive, behavioral, and affective factors. Insight provided by predictive models for these tasks (i.e., a better understanding of how prior preparation, disengaged behavior, and affect are associated with seeking human help) is our primary concern in this work. Cognitive factors we consider are related to performance within the first week of a course as a proxy for prior knowledge of the topic and initial effort in the course. Behavioral factors are related to learner disengagement. We detail the features for each prediction task in §5.4.

- What factors predict that a student will be a TDC Super-User? [Prediction Task #1]
- What factors predict that a student will be a TDC Super-User or a TDC User? [Prediction Task #2]
- What factors predict that a particular student login session within CT will be followed by a TDC session? [Prediction Task #3]

For each prediction task, we also consider overall performance metrics for models we describe in §5.3, including accuracy, precision, recall, and AUC to demonstrate the possibility of delivering successful predictive models for these tasks.

We present the tasks in roughly the order of difficulty from easiest to hardest. In the first task, we attempt to distinguish TDC Super-Users from TDC Non-Users, which we *a priori* expect to be an easier task than distinguishing all users of TDC (i.e., the union of the set of TDC Super-Users and TDC Users) from TDC Non-Users. Finally, looking at individual CT login sessions, we seek characteristics of a student's behavior and affect within the CT session itself as well as general characteristics of the student that may predict she is likely to seek out human help.

The predictive models learned for each of these tasks are retrospective (or perhaps descriptive) in the sense that they rely on data aggregated over students' entire usage of Cognitive Tutor in one or both classes for Prediction Tasks #1 and #2 and data from an entire login session for Prediction Task #3. They serve to help direct future studies toward particular factors that might be included in online algorithms or recommendation systems that implement intelligent instructional hand offs (i.e., in real-time, provide a recommendation that it would be conducive to learning for a student to seek out the help of human tutor from TDC, for example, rather than continue to struggle in Cognitive Tutor).

5. METHODS & APPROACH

In this section, we describe our discovery with models approach, using the output of data-driven behavior and affect detectors as input to classifier models to produce predictions for each of our three prediction tasks. We also describe our iterative under-sampling approach to deal with the extent of imbalance present in this dataset.

5.1 Data-Driven Behavior & Affect Detectors

Extensive literature in educational data mining, learning analytics, human computer interaction, and other disciplines focuses on using sensor-free, data-driven approaches for platforms like ITSs to make inferences about student behavior and affect. This literature has produced a wide variety of "detector" models for various behaviors, especially related to disengagement, and affective states for a bevy of learning platforms (e.g., [5-9, 22, 31, 35]).

In this work, we rely on detectors of disengaged behavior and affect while students use CT. Detectors were implemented for gaming the system [7], off-task behavior [9], and affective states including: boredom, confusion, frustration and engaged concentration [8]. In addition, we implemented contextual models of guessing and slipping to estimate the extent to which each may have been responsible for correct and incorrect answers (i.e., estimating when it may be likely that students are guessing correctly without KC mastery and slipping to produce an incorrect answer despite mastery of a KC) [5]. Contextual slip models have been used as detectors of carelessness in previous work [19, 31]. Gaming the system [6] refers to behavior directed at making progress through content without genuine learning. Learners may try to make progress by adopting strategies like relying on "bottom out" hints that provide the answer or by providing numbers that appear within problem statements as answers to questions, among other shallow (at best) learning strategies.

Detectors we deploy in this study have been successfully used with a similar population of learners in previous work [18-19]. Detectors of gaming the system, off-task behavior, and models of contextual guessing and slipping produce predictions at the level of individual learner actions (i.e., attempts at problem-solving-steps) while detectors of affective states produce predictions about "clips" or time intervals of approximately 20 seconds. For a more extensive summary of the features that are "distilled" from CT log data to serve as input to the underlying machine learning models that constitute these detectors, please see papers cited for each detector [7-9] as well as the papers describing their use with a similar population of higher education CT learners [18-19].

5.2 Imbalanced Data & General Approach

For each prediction task, we adopt an iterative scheme to deal with the fact that each task involves imbalanced data in terms of the target of predictive interest. While a variety of approaches are amenable to the task of dealing with imbalanced data, in the present study, we are primarily interested in establishing the characteristics of disengaged behaviors, affective states, and prior knowledge that predict that students seek out human help, so we adopt a strategy of iteratively considering balanced samples of data, building classifier models on these balanced samples, and considering the factors that contribute to the success of these classifiers. For Prediction Task #1, there are 350 TDC Super-Users and 1,874 TDC Non-Users. For Prediction Task #2, there are 3,119 TDC Super-Users and TDC Users and 1,874 TDC Non-Users. For Prediction Task #3, 3,058 of the 3,119 TDC Super-Users and TDC Users have at least one CT login session that is followed by a session with a TDC tutor while there are 580,528 CT login sessions overall.[4]

For each prediction task, we create a balanced sample by under-sampling the appropriate majority class in each of 500 iterations, building classifier models in each. For Prediction Task #1, this means creating (500x) a sample (with student-level features we describe in §5.4) of the same 350 TDC Super-Users and a random sample of 350 TDC Non-Users. For Prediction Task #2, we create a sample (again, 500x, with student-level features) containing the

[4] 61 students use TDC one or more times before using CT in the courses, so there are no CT sessions from which data can used to better understand what predicts that student's decision to use a TDC tutor.

same 1,874 TDC Non-Users and a random sample of 1,874 students drawn from TDC Users and TDC Super-Users. For Prediction Task #3, we randomly sample one CT session per student that is followed by a TDC session and randomly sample one CT session per student (also chosen at random) that is not followed by a TDC session, resulting in a sample of 6,116 CT sessions for which we have CT session-level features we describe in §5.4. This approach for Prediction Task #3 avoids violations of independence that would be introduced by students with multiple TDC sessions were we to consider more than one session per student.

In each iteration, we have a balanced dataset of student-level or CT-student-session-level features that can be used as predictors in classifier models. We take a 60%-40% split of this dataset into training and test sets, and build classifier models using 5-fold cross validation on the training set, which, given the way we have constructed the training and test set, is student-stratified cross validation. We apply the best performing model in terms of accuracy over this 5-fold cross validation to the held-out test set. Having done this process 500x for each prediction task, we consider the mean (and standard deviation of) performance over these iterations using metrics of accuracy, precision, recall, and area under the ROC curve (AUC). We also consider the specifics of a representative model for each prediction task to provide insights into which features are predictive of seeking out human help.

To test the robustness of this approach, for the case of Prediction Task #2, which is not drastically imbalanced (i.e., 37.5% of students in the sample are non-TDC Users), we consider models learned without using this iterative under-sampling scheme. We show that results are comparable in terms of AUC and compare other performance metrics between the approach, helping to establish possible bounds on expected predictive accuracy and other metrics. Classification accuracy, for example, in this under-sampling scheme is perhaps an especially optimistic estimate of what can be achieved.

5.3 Classifier Models

We consider four types of models to drive classification and prediction: logistic regression (LR), random forest (RF) [13], and support vector machines [16] with both linear (SVML) and radial kernels (SVMR). For each model, we consider the case in which the models output binary classifications as well as probabilities for each of the binary classes of the target variable. In this way, we are able to consider classification accuracy, precision, and recall, as well as AUC as a further comparison of performance compared to chance. Estimated LR models provide a convenient way to consider the significance of features included in these models, so we illustrate the importance of variables in these models in this way.

5.4 Feature Construction

For Prediction Tasks #1 and #2, student-level features are constructed over usage for the entire period of time over the courses in which each student had usage (either the first course, second course, or both). Such features provide for a general profile of how students worked through content in these two courses. Features represent predictions made by detector models as previously described as well as variables related to student performance and usage in their first week of CT usage in the first course they encountered (if they used CT in both courses). Features constructed from "Week 1" data are proxies, however noisy, for student prior preparation and initial knowledge, as other measures, as previously noted, were unavailable. Each variable is constructed as a normalized z-score over all students in the dataset (i.e., the unit for each variable is the number of standard deviations above or below the mean value for each feature):

- *Assistance Per Step*: Mean number of hints requested + errors per problem-solving step
- *Gaming the System*: Proportion of student actions inferred to be instances of gaming the system behavior.
- *Off-Task*: Proportion of student actions inferred to be instances of off-task behavior.
- *Guessing*: Proportion of correct student actions inferred to be possible instances of having correctly guessed.
- *Slipping*: Proportion of incorrect student actions inferred to be possible instances of having slipped despite KC mastery.
- *Boredom*: Proportion of problem solving clips in which students were judged by detector models to have been bored.
- *Frustration*: Proportion of problem solving clips in which students were judged by detector models to have been frustrated.
- *Confusion*: Proportion of problem solving clips in which students were judged by detector models to have been confused.
- *Engaged Concentration*: Proportion of problem solving clips in which students were judged by detector models to be in a state of engaged concentration.
- *Week 1 Sections*: Number of sections of content encountered in the first week of either course (or across both).
- *Week 1 Assistance*: Hints requested and errors made in the first week of either course (or across both).
- *Week 1 Time*: Amount of time spent using Cognitive Tutor in the first week of either course (or across both).
- *Week 1 Sections/Hour*: Number of sections of content encountered per hour in the first week of either course (or across both).
- *Week 1 Assistance/Hour*: Number of hints requested and errors made per hour in the first week of either course (or across both).
- *Week 1 Completer*: binary indicator that a student encountered 90% of the sections in the first week's assignment in either course (or across both).

For Prediction Task #3, CT login-session level features are considered. These features are not normalized, but rather the same proportions as for Prediction Tasks #1 and #2 but with respect to a particular CT login session. For example, Assistance Per Step is calculated over only problem-solving steps within a CT session. *Gaming the System* is calculated as the proportion of student actions within a CT login session that are predicted to be instances of gaming the system, and *Boredom* is calculated as the proportion of problem-solving clips within a CT login session for which detector models infer that a student is bored. Week 1, student-level variables are also included in these models.

6. RESULTS

For each prediction task, we first describe the predictive performance for each of the models we deploy, and then we consider a "representative" logistic regression model that provides insight into the factors that help us to achieve success on these tasks. We describe the sense in which we consider these logistic regression models to be "representative" in the following sub-section.

6.1 Prediction Task 1: TDC Super-Users

We expect the task of distinguishing TDC Super-Users from TDC Non-Users to be the "easiest" *a priori*, in the sense that we expect that we will be able to achieve better performance on the task, an expectation which is borne out by our results. Table 1 shows that logistic regression (LR) performs comparably to a support vector machine with a linear kernel (SVML) with mean accuracy over 500 iterations of .712 and nearly identical values for precision, recall, and AUC. Recall that .5 accuracy represents chance accuracy (and .5 AUC represents chance performance, as ever) because we under-sample to produce a balanced dataset in each iteration.

Table 1. Mean and standard deviation (in parentheses) for accuracy, precision, recall, and area under the ROC curve (AUC) over 500 iterations for the task of predicting whether a student is a TDC Super-User (versus a non-TDC User) [LR = Logistic Regression; RF = Random Forest; SVML = Support Vector Machine with Linear Kernel; SVMR = Support Vector Machine with Radial Kernel]

Model	Accuracy	Precision	Recall	AUC
LR	.712 (.025)	.701 (.029)	.744 (.042)	.786 (.024)
RF	.705 (.025)	.698 (.028)	.727 (.042)	.771 (.024)
SVML	.712 (.024)	.702 (.03)	.743 (.048)	.788 (.023)
SVMR	.665 (.025)	.65 (.03)	.7245 (.056)	.723 (.027)

Table 2 provides a representative, estimated logistic regression model that provides insight into student-level factors that are associated with a student being a TDC Super-User. The model is representative in the sense that, upon inspection of multiple models built on training sets sampled in the way we described above, the significant variables in the model of Table 2 were generally those that were significant. We then specified logistic regression models including only the variables that are reported significant in Table 6 and found that these models, over hundreds of iterations, achieved results nearly identical to those reported for logistic regression in Table 1. Spot inspections of model parameters in numerous models produced by the iterative process also aligned with those reported in Table 2 in terms of both sign and magnitude. This same notion of representative logistic regression models is used for each of the three predictive tasks we consider to provide insight into the variables that contribute to such models.

The model of Table 2 suggests that the four significant factors for predicting that a student will be a TDC Super-User are *Off-Task* disengagement, *Boredom*, *Guessing,* and *Week 1 Sections/Hour*. Pairwise Pearson correlations among these significant predictors are small, with no statistically significant correlation between *Guessing* and *Off-Task* disengagement, and the largest significant correlation is that between *Week 1 Sections/Hour* and *Boredom* ($r = .36$; $p < .001$). These observations, combined with the consistency of models learned over only these significant predictors, instill confidence in our interpretation of the logistic regression coefficients. However, multi-collinearity among some of the other predictors (especially, for example, *Gaming the System* and *Confusion*: $r = .76$; $p < .001$) requires us to exercise caution in interpreting other coefficients in this representative logistic regression model. Roughly these same observations about the significant predictors as well as caveats concerning the interpretation of the non-significant estimated regression coefficients are operative for Predictive Tasks #2 and #3.

While disengagement is positively associated with TDC Super-User status, the *Boredom*, *Guessing*, and *Week 1 Sections/Hour* are negatively associated with TDC Super-User status, indicating that students who are inferred to be less bored, less likely to be haphazardly guessing, and better prepared for the coursework (as indicated by efficient progress through content in the first week of the course) are less likely to seek out human help extensively.

Table 2. Representative estimated logistic regression model for the task of predicting whether a student is a TDC Super-User (versus a non-TDC User). *Rows for significant variables at α = 0.05 are bold and italicized.*

Variable	Coefficient	Std. Error	p-value
(Intercept)	-.756	.604	.21
Assistance Per Step	.664	.526	.207
Gaming the System	.103	.271	.704
Off-Task	***.35***	***.161***	***.03***
Guessing	***-.611***	***.306***	***.046***
Slipping	.135	.17	.429
Boredom	***-.774***	***.292***	***.009***
Frustration	.249	.182	.172
Confusion	-.084	.233	.72
Engaged Concentration	.376	.306	.218
Week 1 Sections	.361	.193	.061
Week 1 Assistance	-.333	.322	.302
Week 1 Time	.058	.277	.833
Week 1 Sections/Hour	***-1.365***	***.425***	***.001***
Week 1 Assistance/Hour	.052	.285	.856
Week 1 Completer	.478	.636	.452

6.2 Prediction Task 2: TDC Users + TDC Super Users

As expected, we find that distinguishing those students who used TDC at least once (the set of TDC Users + TDC Super-Users) from TDC Non-Users is more "difficult" in the sense that models achieve a lower degree of classification accuracy, precision, and recall, as well as a lower AUC (Table 3).

Inspection of the representative, estimated LR model in Table 4 indicates that in addition to the three same features that are significant in predicting TDC Super-User status (i.e., *Off-Task*

disengagement, *Boredom*, and *Guessing*), *Week 1 Time* is a significant predictors that students will have used TDC at least once, suggesting that this measure of time provides different information to help distinguish between these categories of students.

Table 3. Mean and standard deviation (in parentheses) for accuracy, precision, recall, and area under the ROC curve (AUC) over 500 iterations for the task of predicting whether a student used TDC at least once (i.e., TDC Super-User or TDC User versus a non-TDC User) [see model acronyms in caption for Table 1]

Model	Accuracy	Precision	Recall	AUC
LR	.614 (.0111)	.62 (.012)	.592 (.023)	.666 (.012)
RF	.615 (.0109)	.614 (.0115)	.618 (.0209)	.66 (.0113)
SVML	.612 (.0105)	.624 (.0133)	.5651 (.0377)	.6656 (.0113)
SVMR	.598 (.0115)	.6 (.013)	.591 (.0332)	.629 (.0121)

Table 4. Representative estimated logistic regression model for the task of predicting whether a student is a TDC User (versus a non-TDC User). *Rows for significant variables at α = 0.05 are bold and italicized.*

Variable	Coefficient	Std. Error	p-value
(Intercept)	-.1	.2	.617
Assistance Per Step	-.068	.118	.564
Gaming the System	-.05	.092	.586
Off-Task	***.133***	***.063***	***.035***
Guessing	***-.233***	***.071***	***<.001***
Slipping	-.021	.056	.706
Boredom	***-.527***	***.088***	***<.001***
Frustration	.067	.045	.142
Confusion	.132	.092	.149
Engaged Concentration	.042	.096	.661
Week 1 Sections	-.117	.064	.067
Week 1 Assistance	-.226	.115	.05
Week 1 Time	***.378***	***.128***	***.003***
Week 1 Sections/Hour	-.135	.077	.08
Week 1 Assistance/Hour	-.085	.081	.29
Week 1 Completer	.205	.213	.335

Since this prediction task is the least imbalanced of the three we consider, we also consider learning models without our adopted under-sampling scheme. Though we omit extensive analysis of these models for brevity, Table 5 provides performance metrics for LR and RF models learned by taking a 60-40% training-test split of all students, learning models using 10-fold cross validation on the training set and applying the model with greatest accuracy to the test set. We find that this model modestly out-performs the trivial, majority class classifier in terms of classification accuracy with comparable precision, but recall of this model is substantially greater than that achieved by typical models in our under-sampling scheme.

Building on our observations from the previous model, since *Week 1 Time* has a positive parameter estimate, students who take more time to work through content in the first week, and perhaps work more diligently by guessing less as they make problem-solving attempts, may be more likely to seek out help via TDC. It is possible that otherwise relatively diligent students (by some measures) who seek out TDC begin to adopt a sub-optimal learning strategy of some sort that is indicated by the *Off-Task* detector more frequently than those students who do not seek out TDC.

Consequently, the F measure (one commonly used evaluation metric that balances precision and recall) would be greater for these models than for those of the typical models of our under-sampling scheme. Nevertheless, AUC of these models are nearly identical to mean values of models learned according to our under-sampling scheme. Perhaps more importantly, the estimated logistic regression model points to exactly the same set of significant behavioral and affective features, *Off-Task* disengagement, *Boredom*, and *Guessing*, as the model reported in Table 4. *Week 1 Time* is also significant in models using both approaches, though *Week 1 Sections*, *Week 1 Assistance/Hour*, and *Week 1 Completer* are significant in the model that does not rely on under-sampling.

Table 5. Accuracy, precision, recall, and area under the ROC curve (AUC) for the task of predicting whether a student used TDC at least once (versus a TDC Non-User) for models estimated without relying on under-sampling scheme (trivial majority classifier accuracy = .625)

Model	Accuracy	Precision	Recall	AUC
LR	.666	.684	.865	.669
RF	.651	.68	.832	.659

6.3 Prediction Task 3: TDC Sessions Follows a CT Login Session

As expected, the most difficult task was to predict whether a particular CT session was going to be followed by a TDC session, as illustrated by the performance metrics for the various models we consider in Table 6.

Table 6. Mean and standard deviation (in parentheses) for accuracy, precision, recall, and area under the ROC curve (AUC) over 500 iterations for the task of predicting whether a particular student CT session is followed by a session with a TDC tutor [see model acronyms in caption for Table 1]

Model	Accuracy	Precision	Recall	AUC
LR	.599 (.009)	.604 (.015)	.579 (.021)	.633 (.01)
RF	.587 (.01)	.587 (.014)	.592 (.023)	.621 (.011)
SVML	0.6 (.009)	.61 (.015)	.559 (.023)	.633 (.01)
SVMR	.598 (.01)	.606 (.017)	.567 (.031)	.632 (.01)

Table 7 provides a representative, estimated LR model that provides insight into the factors that are predictive of a student's tendency to seek out human tutoring via TDC from within a particular CT session. Here, *Boredom* appears again, along with *Engaged Concentration* (which was significant in neither Prediction Task #1 nor Prediction Task #2), as a significant, negatively associated predictor. We also find that *Gaming the System*, another form of disengagement, is positively associated with a tendency to seek out immediate help via TDC, along with *Week 1 Sections*.

At the level of student-login sessions in Prediction Task #3, *Gaming the System* and the other detected factors are no longer highly correlated (as they were when we considered student-level aggregated features in Prediction Tasks #1 and #2). Rather # *Hints* and # *Errors* and *Week 1 Time* and *Week 1 Assistance* are relatively highly correlated, leading us to exercise caution in the interpretation of estimated coefficients associated with these (insignificant) predictors.

Table 7. Representative estimated logistic regression model for the task of predicting whether a particular student CT login session is followed by a session with a TDC tutor. Coefficients are un-standardized. *Rows for significant variables at α = 0.05 are bold and italicized.*

Variable	Coefficient	Std. Error	p-value
(Intercept)	***1.321***	***.364***	***<.001***
# Errors	-.002	.001	.177
# Hints	.002	.001	.262
Gaming the System	***1.367***	***.23***	***<.001***
Off-Task	.612	.652	.348
Guessing	-.718	1.207	.552
Slipping	-.197	.491	.689
Boredom	***-.783***	***.149***	***<.001***
Frustration	.12	.316	.705
Confusion	.213	1.32	.872
Engaged Concentration	***-1.575***	***.229***	***<.001***
Week 1 Sections	***.021***	***.006***	***<.001***
Week 1 Assistance	-.0001	.0001	.179
Week 1 Time	.003	.01	.733
Week 1 Sections/Hour	-.021	.019	.263
Week 1 Assistance/Hour	-.0004	.001	.68

7. DISCUSSION

7.1 Highlights & Summary

At least two qualitative findings are robust in the modeling presented. First, as inferred by detector models in CT, *Boredom* is negatively associated with a tendency to seek out human help outside of the CT ITS via the TDC service in both the aggregate (Prediction Tasks #1 and #2) as well as the more immediate term (Prediction Task #3). Especially when combined with the negative association of *Guessing* with seeking out TDC's services in Prediction Tasks #1 and #2, this suggests at least a modicum of baseline diligence in working within CT for those who sought out TDC. However, the second robust finding may point to the adoption of counter-productive strategies that may also lead students to require assistance outside of the ITS. This second robust finding is that two facets of learner disengagement inferred by such detector models, *Off-Task* behavior and *Gaming the System*, are positively associated, in the aggregate and more immediately, respectively, with seeking human assistance outside of the CT ITS. These insights contribute to a bevy of literature concerning various aspects of the technology-enhanced learning experience, generally centered on learning outcomes and learners using ITSs, which are associated with these phenomena (e.g., [14, 18, 30]).

7.2 Limitations

While we consider a rich, substantial data set with thousands of learners, the present analysis is not without its limitations. First, we merely consider learning models to predict that a student is likely to be particular "type" of TDC user or that a particular CT login session is likely to be followed by a session with a TDC tutor. We do not consider the effectiveness of TDC sessions, though some work has begun to consider that question [28-29], or attempt to deeply link the specific KCs within CT on which students may have been working when they sought out TDC. This dataset also offers the opportunity to consider CT usage and performance (possibly at the level of fine-grained KCs) before and after a TDC session as a type of pre- and post-test for these sessions.

Further, this is a purely retrospective, observational study, and the empirical frequencies with which students sought out (and did not seek out) help via the TDC service reflects likely over-use and near certain under-use. While the models we have learned have provided insights into the context in which these data were collected, data from scenarios and contexts in which we suspect that such use of human tutors is more attuned to need would provide interesting contrast cases to the present study. In addition, while associations uncovered by predictive models like those presented could arise due to causal relationships between factors captured by these predictors, the present analysis does not provide us evidence for any such claims. While adopting counter-productive strategies like gaming the system in CT may precede seeking out human help, is such a counter-productive strategy really the cause of seeking such help? If we were to conceive of a clever intervention to reduce gaming the system behavior, would that reduce the incidence of learners seeking out human help? Future work might more carefully observe students in environments in which they can seek out human help while using an ITS (or other systems) to elicit their explanations for help seeking, or experimental studies might consider interventions that tend to increase or decrease the extent to which students rely on external help.

8. FUTURE WORK

In addition to several opportunities noted in the previous section, we consider two "big" ideas with respect to future work.

8.1 Information vs. Affirmation

One concern with this analysis is that we are building models that combine different motives for students to seek out human

assistance. Consider the following dialog (a slightly edited TDC interaction):

Tutor: hi! what can I help you with today?

Student: Do you know how to do a factor table?

Tutor: Hmm I am familiar with it. Is there a problem that you wanted to go over?

Student This looks like an easy one, but I am not sure so I just want to make sure I understand this correctly

Student: To check this table is all you do multiply the top row by the 7x and see if it matches the bottom row? Is this right?

Tutor: Yeah everything looks good to me**.** Great job!

Student: I was hoping that I did this right.

We call this kind of interaction a request for "affirmation," rather than information. The tutor is not teaching the student anything, just verifying that the student's approach is correct. The conditions leading to this type of interaction are likely to be very different from information requests. They may occur when students have high knowledge but low confidence, for example. Future work will explore models that separate information from affirmation sessions.

8.2 Instructional Hand Offs

In contemporary K-12 classrooms, online courses, and other settings for learning, students may seek instruction, assistance, remediation, opportunities for enrichment, and even affirmation from multiple resources, including technology resources like ITSs and non-technological resources like human beings. Especially when at least one of these resources is technological, providing adaptive, intelligent guidance to learners as to when they should use particular resources and applications (or persist and try to "stick with it" and learn within a particular application) will be crucial. In the present study, we have sought to better understand cognitive, behavioral, and affective factors that predict that a student may seek help from a non-ITS resource like a human tutor while using the CT ITS for math, but other types of instructional hand offs should also be considered.

Hand offs between instructional applications might happen, for example, between an ITS and a simulation-based training environment. When a student has completed all of the skills for which the ITS provides instruction, the simulation-based training environment that includes some overlapping content with the ITS could tailor its simulated scenarios around emphasizing elements of those skills in the ITS on which the student struggled and de-emphasize skills that the student easily mastered within the ITS. This is likely to require a *lingua franca* shared by the ITS and the simulator about the competencies or skills that are tracked by each, or perhaps both may rely on a set of external standards or some other way of indicating how this type of hand off based on such cognitive factors may work. Efforts including the development of the Experience API[5] (xAPI), the Total Learning Architecture[6] (TLA), and the Generalized Intelligent Framework for Tutoring[7] (GIFT) exemplify moves in directions that would enable progress toward these and similar goals.

[5] https://github.com/adlnet/xAPI-Spec

[6] https://www.adlnet.gov/tla/

[7] https://www.gifttutoring.org/

Of course, even in the case of guiding an instructional hand off between an ITS and a human tutor (or K-12 classroom teacher) for a student who needs help with content covered by the ITS, the ITS ideally should be able to communicate to the human tutor or classroom teacher that the learner in question requires assistance on a particular skill, just needs a confidence boost, or has been adopting counter-productive and/or disengaged learning strategies like gaming the system that should probably be discouraged. Insights into predictors of help seeking may help to drive development of recommendations delivered by the learning application to the learner or could also drive recommendations to a teacher via an application that surfaces insights from the ITS.

We hope this work provides a step toward more work on these, and related, problems.

9. ACKNOWLEDGMENTS

This work was funded by a contract from the U.S. Department of Defense Advanced Distributed Learning Initiative (Contract PAAIDT W911QY-14-C-0019). Anonymous reviewers provided helpful comments that have improved the presentation of this work.

10. REFERENCES

[1] Aleven, V., and Koedinger, K. R. 2000. Limitations of student control: Do students know when they need help? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, (Montreal, Canada, June 19-23, 2000). ITS 2000. Springer-Verlag, Berlin, 292-303.

[2] Aleven, V., McLaren, B., Roll, I., and Koedinger, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence and Education* 16, 2, 101-128.

[3] Aleven, V. Stahl, E., Schworm, S. Fischer, F. and Wallace, R. 2003. Help seeking and help design in interactive learning environments. *Rev. Educ. Res.* 73, 3, 277-320.

[4] Almeda, V., Baker, R., and Corbett, A. 2017. Help avoidance: When students should seek help, and the consequences of failing to do so. *Teach. Coll. Rec.* 117, 3, 1-24.

[5] Baker, R.S., Corbett, A.T., and Aleven, V. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems,* (Montreal, Canada, 2008). ITS 2008. 406-415

[6] Baker, R.S., Corbett, A.T., Koedinger, K.R., and Wagner, A.Z. 2004. Off-task behavior in the Cognitive Tutor classroom: when students "game the system." In *Proceedings of ACM CHI 2004: Computer-Human Interaction* (Vienna, Austria, 2004). 383-390.

[7] Baker, R.S., and de Carvalho, A. M. J. A. 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (Montreal, 2008). 38-47.

[8] Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., and Rossi, L. 2012. Towards sensor-free affect detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, (Chania, Greece, 2012). 126-133.

[9] Baker, R.S. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction* (San Jose, CA, April 28 – May 3, 2007). ACM, New York, 1059-1068.

[10] Baker, R.S., and Yacef, K. 2009. The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining* 1, 3-17.

[11] Bloom, B. S. 1968. Learning for mastery. *Evaluation Comment* 1, 2, 1–12.

[12] Bloom, B. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Researcher* 13, 6, 4-16.

[13] Breiman, L. 2001. Random forests. *Mach. Learn.* 45, 1, 5-32.

[14] Cocea, M., Hershkovitz, A., Baker, R.S.J.d. 2009. The impact of off-task and gaming behavior on learning: immediate or aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (Brighton, UK, 2009). 507-514

[15] Corbett, A.T., and Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adap.* 4, 253-278.

[16] Cortes, C., and Vapnik, V.N. 1995. Support-vector networks. *Mach. Learn.* 20, 3, 273–297.

[17] Daniels, C., Edwards, R., Miller, P., Hale, A., Powell, L., Wisner, J., Mallonee, K., Perkins, S., Bravo, J., Hummel, M., Wagner, B., and Fay, M. 2010. *PowerTeaching: Cooperative Learning Handbook.* Success For All Foundation, Baltimore, MD.

[18] Fancsali, S.E. 2014. Causal discovery with models: behavior, affect, and learning in Cognitive Tutor Algebra. In *Proceedings of the 7th International Conference on Educational Data Mining,* (London, UK, 2014). EDM 2014. 28-35.

[19] Fancsali, S.E. 2015. Confounding carelessness? Exploring causal relationships between carelessness, affect, behavior, and learning in Cognitive Tutor Algebra. In Proceedings of the 8th International Conference on Educational Data Mining, (Madrid, Spain, 2015). EDM 2015. 508-511.

[20] Fancsali, S.E., Ritter, S., Berman, S.R., Yudelson, M., Rus, V., and Morrison, D.M. 2016. Toward integrating Cognitive Tutor interaction data with human tutoring text dialogue data in LearnSphere. In *Proceedings of the EDM 2016 Workshops and Tutorials*, (Raleigh, NC, 2016). CEUR Workshop Proceedings.

[21] Fancsali, S.E., Rus, V., Ritter, S., and Berman, S.R. 2017. *Final Technical Report: Integrating Human and Automated Tutoring Systems*. Technical Report. Carnegie Learning, Inc., Pittsburgh, PA.

[22] Johns, J. and Woolf, B. 2006. A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the 21st National Conference on Artificial Intelligence* (Boston, MA, 2006). AAAI Press, Menlo Park, CA, 2-8.

[23] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J. 2011. A data repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker, Eds. CRC, Boca Raton, FL, 43-55.

[24] Koedinger, K.R., Corbett, A.T., and Perfetti, C. 2012. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Sci.* 36, 757-798.

[25] Renkl, A. 2002. Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learn. Instr.* 12, 529-556.

[26] Ritter, S., Anderson, J.R., Koedinger, K.R., and Corbett, A.T. 2007. Cognitive Tutor: applied research in mathematics education. *Psychon. B. Rev.* 14, 249-255.

[27] Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R. 2011. Improving students' help seeking skills using metacognitive feedback in an intelligent tutoring system. *Learn. Instr.* 21, 267-280.

[28] Rus, V., Banjade, R., Maharjan, N., Morrison, D., Ritter, S., and Yudelson, M. 2016. Preliminary results on dialogue act classification in chatbased online tutorial dialogues. In *Proceedings of the 9th International Conference on Educational Data Mining*, (Raleigh, NC, 2016). EDM 2016. 630-631.

[29] Rus, V., Maharjan, N., Tamang, L.J., Yudelson, M., Berman, S., Fancsali, S.E., and Ritter, S. 2017. An analysis of human tutors' actions in tutorial dialogues. In *Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference*, (Marco Island, FL, 2017). FLAIRS-30. 122-127.

[30] San Pedro, M.O.C.Z., Baker, R.S., Bowers, A.J., and Heffernan, N.T. 2013. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Proceedings of the 6th International Conference on Educational Data Mining* (Memphis, TN). EDM 2013. 177-184

[31] San Pedro, M.O.C.Z., Baker, R. S., Rodrigo, M. 2011. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education* (Auckland, New Zealand). 304-311.

[32] Searle, J.R. (1969). *Speech Acts*. Cambridge UP.

[33] Sosnovsky, S. 2011. Ontology-Based Open-Corpus Personalization for E-Learning. Doctoral Thesis. University of Pittsburgh.

[34] vanLehn, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* 6, 4, 197-221.

[35] Walonoski, J., and Heffernan, N.T. 2006. Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, (Jhongli, Taiwan, 2006). ITS 2006. 722-724.

[36] Wood, H., and Wood, D. 1999. Help seeking, learning and contingent tutoring. *Comput. Educ.* 33, 153-169.

[37] Zimmerman, B. J. 1990. Self-regulated learning and academic achievement: An overview. *Educ. Psychol.* 25, 1, 3-17.